Statistical assessment of chromosomal aberrations at the cohort level: the CGHSeg package

Franck Picard, Baba Thiam, Stephane Robin

CGHSeg
Statistical assessment of chromosomal aberrations at the cohort level

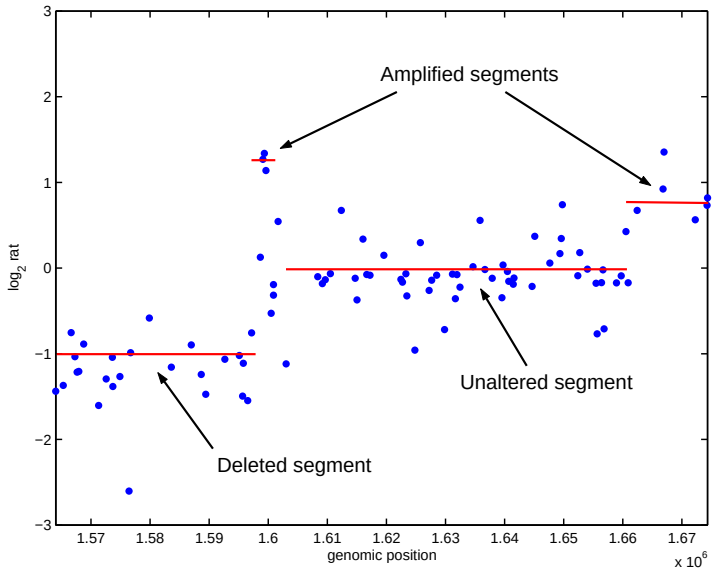
F. Picard[‡,⋆,◇], M. Hoebeke[⋆], E. Lebarbier[†], B. Thiam[†], S. Robin[†]

[‡]UMR 5558 UCB CNRS LBBE, Lyon, France
[◇] Projet BAMBOO, INRIA, F-38330 Montbonnot Saint-Martin, France.
[†] UMR 518 AgroParisTech/INRA, F-75231, Paris, France
[⋆] UMR CNRS 8071-INRA 1152-UEVE F-91000 Evry, France

Rennes, July 8th 2009

The basics of aCGH

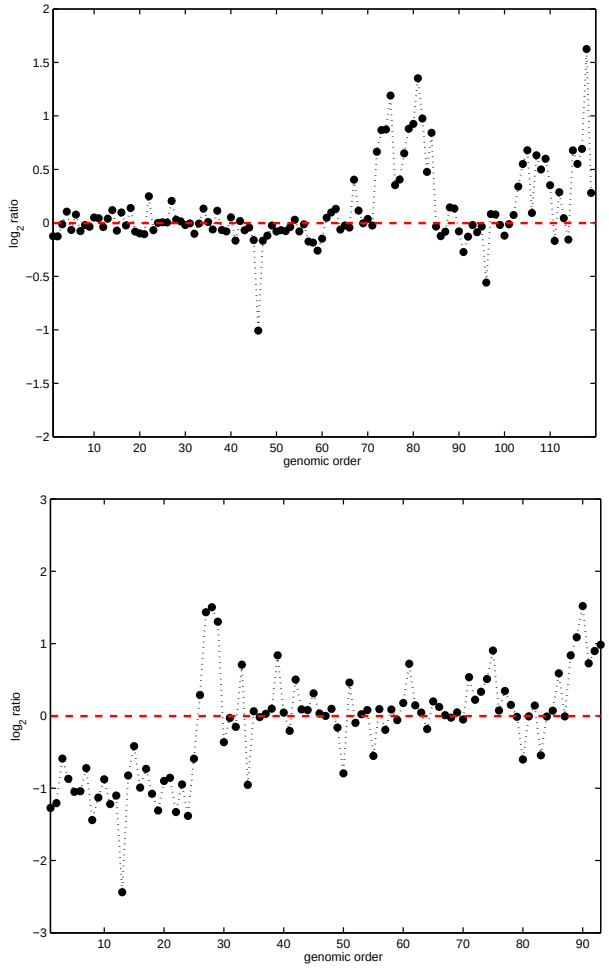
- Investigation of Chromosomal aberrations
- At the genome scale
- Using the microarray technology



$$\log_2 \left\{ \frac{\sharp \text{ copies of BAC(t) in the test genome}}{\sharp \text{ copies of BAC(t) in the reference genome}} \right\}$$

First years of array CGH data analysis

- **First papers**:
  2002 Olshen et al.
  2004 Fridlyand et al. Hupé et al.
  2005 Picard et al.
- **Motivations**:
  find breakpoints
  assign a status to segments
- **Frameworks**:
  segmentation HMMs smoothing.



The CGHSeg package

- Segmentation for aCGH,
- uni-patients and multi-patients,
- Uses C++ and S4 classes.
  - → CGHdata
  - → CGHoptions
  - → CGHresults

```
****** Summary of CGHd object ******
[CGHd summary] Chromosomes id : 8
[CGHd summary] Groups id      : 1 2 3 4
[CGHd summary] Patients per group
[CGHd summary] Group  1  :  11  patients
X309 X387 X503 X504 X509 X517 X519 X549
X571 X574 X98
...
[CGHd summary] recorded variables
             class
group       factor
patient     factor
chromosome  factor
phys.pos   numeric
order       factor
signal     numeric
clone.id    factor
age        numeric
sex         factor
location    factor
```

Definitions and notations for segmentation models

- We observe $\mathbf{Y} = \{Y_1, \ldots, Y_n\}$ (i.i.d.) :

$$Y_t \sim \mathcal{N}(\mu_t, \sigma^2).$$

- We suppose that there exists breakpoints $\mathbf{T} = \{t_1, \ldots, t_K\}$:

$$\forall t \in I_k, \quad Y_t = \mu_k + E_t, \quad E_t \sim \mathcal{N}(0, \sigma^2)$$

- $\boldsymbol{\mu}$ corresponds to the mean of segments,
- $\mathbf{T}$ corresponds to the breakpoint positions.

CGHSeg for uni-patient segmentation

- Get $(\hat{\mathbf{T}}, \hat{\mu})$ by Dynamic Programming
- `unisegmean`: segmentation in the mean[1]
- `unisegclust`: segmentation/clustering[2]
- Model selection: `adaptive`[1], `mBIC`[3]

```
> CGHo = new("CGHoptions")
> CGHo
****** CGHoption show ******
     options value
1     select  adaptive
2      clust  FALSE
3  poseffect  TRUE
4       Pmin     2
5       Pmax     5
6       lmin     1
7       lmax     1
8      alpha   0.1
9       beta   0.1
10      fast FALSE
11    output   all
> CGHr = uniseg(CGHd,CGHo)

> clust(CGHo) = TRUE
> CGHr        = uniseg(CGHd,CGHo)
```

Multiple samples analysis

- Chromosomal aberrations
  - (*i*) can be used for efficient tumor classification,
  - (*ii*) are associated with overall survival of patients,
  - (*iii*) are linked to differential response to various cancer therapies.
- Study of multisamples with the same plateform,
- The purpose is the joint characterization of their CGH profiles,
- They share technical bias (probe effect, 'wave effect').

Joint segmentation of multi-patient profils

- We now observe Y_t^m, the signal for patient m at position t
- There exists a probe effect which is common to all patients
- The mean of Y_t^m is still subject to changes:

$$\forall t \in I_k^m,\ Y_t^m = \mu_k^m + \theta_t + \varepsilon_t^m \text{ with } \varepsilon_t^m \sim \mathcal{N}(0, \sigma^2)$$

- $\boldsymbol{\theta}$ will be used for normalization purposes
- Get $(\hat{\mathbf{T}}, \hat{\boldsymbol{\mu}})$ by Dynamic programming
- Get $\hat{\boldsymbol{\theta}}$ by Least Squares $\longrightarrow$ `ILS()` functions (Iterative LS)

Joint segmentation/clustering of multi-patient profils

- The mean of the signal should be restricted to $\{m_1, \ldots, m_P\}$,
- We $\{Z^k = P\}$ the label of segment k
- Given $\{Z^k = P\}$:

$$\forall t \in I_k^m, \; Y_t^m = m_P + \theta_t + \varepsilon_t^m \text{ with } \varepsilon_t^m \sim \mathcal{N}(0, \sigma^2)$$

- Get $(\hat{\mathbf{T}})$ by Dynamic programming
- $\hat{\mathbf{m}}$ by the EM algorithm,
- Get $\hat{\boldsymbol{\theta}}$ by Least Squares $\longrightarrow$ `ILSclust()` functions

A 2-stage Dynamic Programming

- Minimize the RSS:

$$RSS_K(\boldsymbol{\mu}, \mathbf{T}) = \sum_{m=1}^{M} \sum_{k=1}^{K_m} RSS_k^m(\boldsymbol{\mu}_m, \mathbf{T}_m) = \sum_{m=1}^{M} \sum_{k=1}^{K_m} \sum_{t \in I_k^m} (y_{mt} - \mu_{km})^2,$$

- But there is a constraint : $\sum_m K_m = K$, thus:

$$\min_{\{\mathbf{T}, \boldsymbol{\mu}\}} RSS_K(\mathbf{T}, \boldsymbol{\mu}) \quad = \quad \min_{K_1 + \ldots + K_M = K} \left\{ \sum_{m=1}^{M} \min_{\mathbf{T}_m, \boldsymbol{\mu}_m} RSS_{K_m}^m(\mathbf{T}_m, \boldsymbol{\mu}_m) \right\}$$

CGHSeg for multi-patient segmentation

- Get $(\hat{\mathbf{T}}, \hat{\mu})$ by 2-stage DP
- Underlying functions of `multiseg()`
  - → with correction: `ILS()`, `ILSclust()`
  - → without correction: `multisegmean()` `multisegclust()`

```
> CGHr = multiseg(CGHd,CGHo)
[multiseg] ILS running

> CGHr["mu"][['chr8']][['group1']][['X607']]
  begin end         mean
1     1  23 -0.459185095
2    24  72 -0.003737113
3    73 137  0.282555851
...
> CGHr["theta"]
$`chr8`
  [1] -0.145 -0.031  0.014 -0.128 -0.035...
```

CGHSeg for multi-patient segmentation

- Get $(\hat{\mathbf{T}}, \hat{\mu})$ by 2-stage DP
- Underlying functions of `multiseg()`
  - → with correction: `ILS()`, `ILSclust()`
  - → without correction: `multisegmean()` `multisegclust()`

```
> CGHo        = new("CGHoptions")
> clust(CGHo) = TRUE
> CGHr        = multiseg(CGHd,CGHo)
[multiseg] ILSclust running

> CGHr["mu"][['chr8']][['group1']][['X585']]
  begin end         mean clust
1     1  43 -0.009450802     1
2    44  50  0.451431544     2
3    51  64 -0.009450802     1
4    65 137  0.451431544     2
...
> CGHr["theta"]
$`chr8`
  [1] -0.273 -0.160 -0.118 -0.266 -0.152...
```

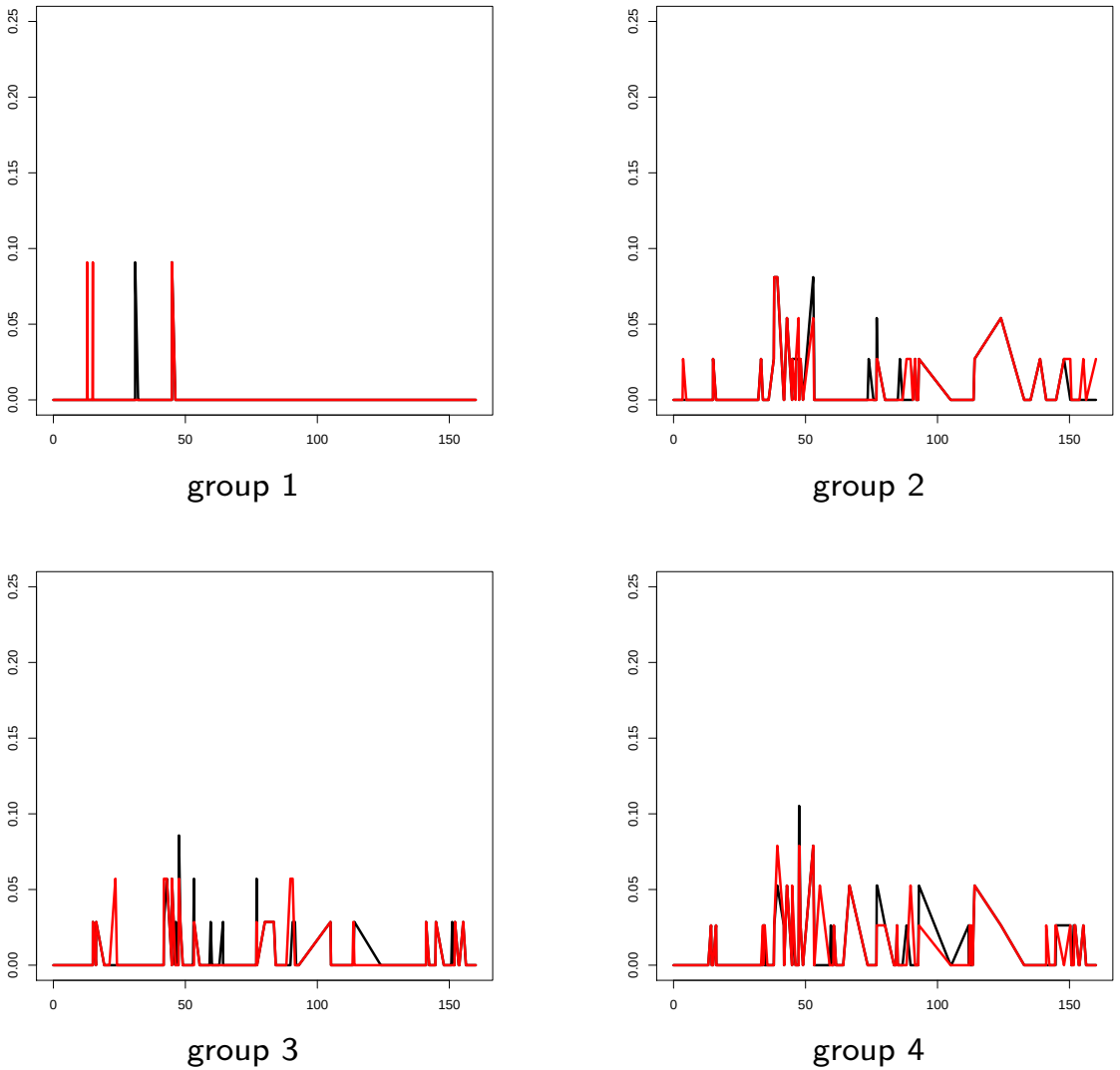
Handling results of `multiseg()` functions

- From `CGHr` we can get many features of the model
- the breakpoints frequencies across patients
- the predictions/residuals for each patient/group
- the clusters frequencies per position

```
> bp(CGHr,CGHo,by = "patient")
> bp(CGHr,CGHo,by = "group")

> resid(CGHr,CGHd,CGHo,by = "patient")
> resid(CGHr,CGHd,CGHo,by = "group")

> predict(CGHr,CGHo,by = "patient")
> predict(CGHr,CGHo,by = "group")

> clusterfreq(CGHr,CGHo)
```

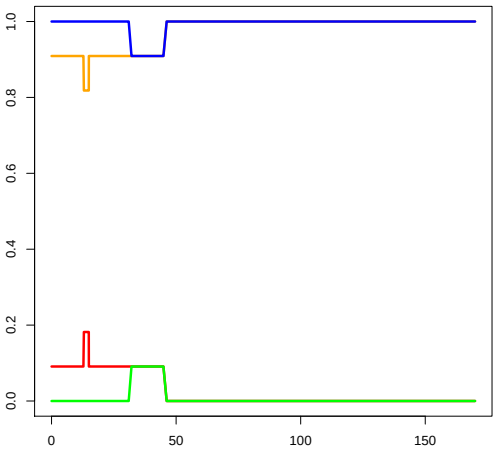
Breakpoint frequencies vs genomic position (Nakao-chr8)

```
bp(CGHr,CGHo,by = "group")
```

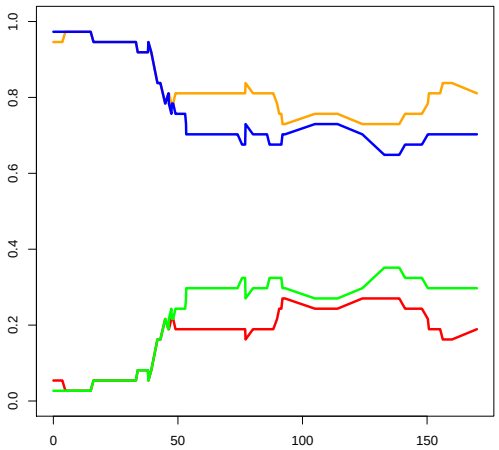


Cluster frequencies vs genomic position (Nakao-chr8)

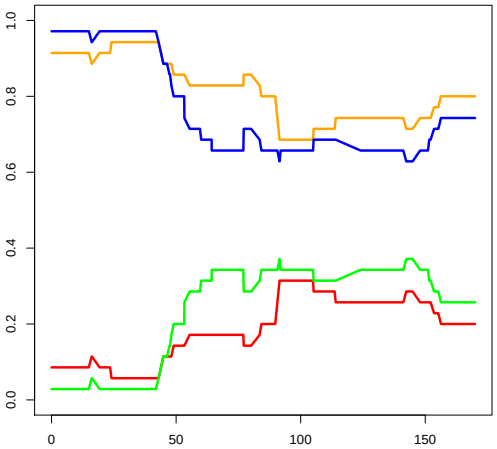
`clusterfreq(CGHr,CGHo)`



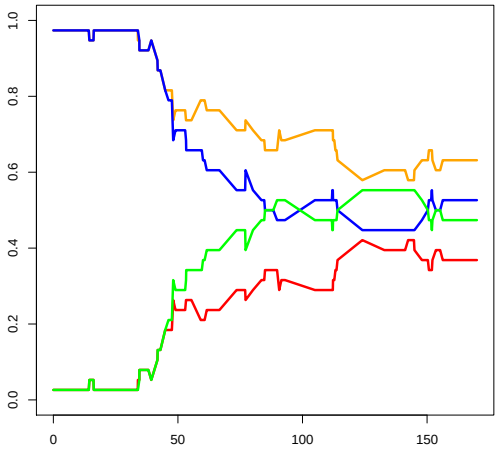
group 1



group 2



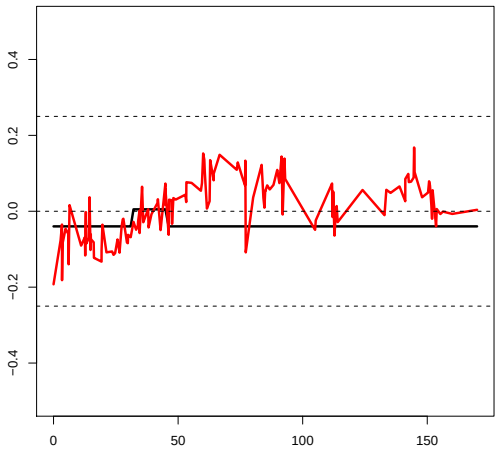
group 3



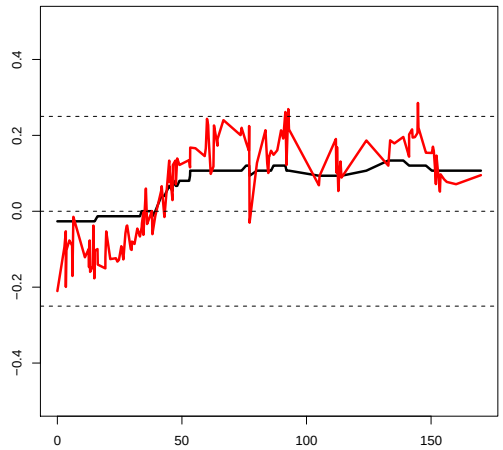
group 4

Mean Profiles vs genomic position (Nakao-chr8)

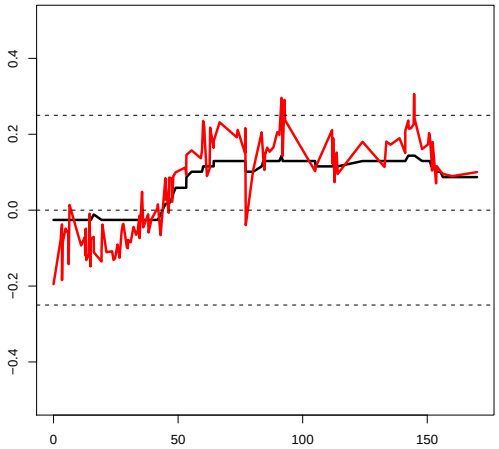
```
predict(CGHr,CGHo,by = "group")
```



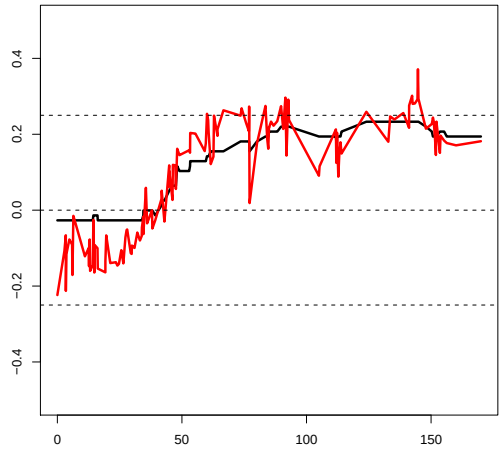
group 1



group 2



group 3



group 4

Conclusions

- The `CGHSeg` is designed for segmentation on array CGH data
- It gather different works on process segmentation and model selection
- Could be extended to add more normalization effects, experimental design
- Soon available on the CRAN

References

F. Picard, S. Robin, M. Lavielle, C. Vaisse, and J.-J. Daudin.

A statistical approach for array CGH data analysis.

BMC Bioinformatics, 6(27):1, 2005.

F. Picard, S. Robin, E. Lebarbier, and J.-J. Daudin.

A segmentation/clustering model for the analysis of array cgh data.

Biometrics, 63:758–756, 2007.

N.R. Zhang and D.O. Siegmund.

A modified bayes information criterion with applications to the analysis of comparative genomic hybridization data.

Biometrics, 63(1):22–32, 2007.